DOCUMENT RESUME

ED 093 928 TM 003 746

AUTHOR Kane, Michael T.; Moloney, James M.
TITLE The Effect of SSM Grading on Reliability When Residual Items Have No Discriminating Power.
PUB DATE [Apr 74]
NOTE 5p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Feedback; Guessing (Tests); *Multiple Choice Tests; *Response Style (Tests); *Scoring Formulas; Statistical Analysis; *Test Reliability
IDENTIFIERS *Self Scoring Method; SSM

ABSTRACT

Gilman and Ferry have shown that when the student's score on a multiple choice test is the total number of responses necessary to get all items correct, substantial increases in reliability can occur. In contrast, similar procedures giving partial credit on multiple choice items have resulted in relatively small gains in reliability. The analysis in this paper provides a possible explanation for this discrepancy. The reliability for SSM (self-scoring method) grading is compared to the reliability for conventional 0-1 grading by postulating a model for the conditional distributions of a correct response on the n-th try for an item, given an incorrect response on the first try. The results imply that for SSM grading, distractors should have different levels of attractiveness rather than being equally attractive. (Author)

ED 093928

TM 003 243

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE EFFECT OF SSM GRADING ON RELIABILITY WHEN RESIDUAL ITEMS HAVE NO DISCRIMINATING POWER

Michael T. Kane

SUNY, Stony Brook

and

James M. Moloney

SUNY, Brockport

Abstract

Gilman and Ferry have shown that when the student's score on a multiple choice test is the total number of responses necessary to get all items correct, substantial increases in reliability can occur. In contrast, similar procedures giving partial credit on multiple choice items have resulted in relatively small gains in reliability. The analysis in this paper provides a possible explanation for this discrepancy. The reliability for SSM grading is compared to the reliability for conventional 0-1 grading by postulating a model for the conditional distributions of a correct response on the n-th try for an item, given an incorrect response on the first try. The results imply that, for SSM grading, distractors should have different levels of attractiveness rather than being equally attractive.

THE EFFECT OF SSM GRADING ON RELIABILITY WHEN RESIDUAL ITEMS HAVE NO DISCRIMINATING POWER

Michael T. Kane, SUNY, Stony Brook
and James M. Moloney, SUNY, Brockport

Introduction

Several alternate scoring procedures have been proposed for multiple choice items with the aim of increasing their discriminating power. Empirical studies of these procedures have generally indicated that, for a given set of items, reliability can be increased by substituting the newer grading procedures for the conventional 0-1 grading system (Hambleton, Roberts, and Traub, 1970). The improvement in reliability is generally small and, in at least one case, the conventional grading system had the higher reliability (Koehler, 1971). Gilman and Ferry (1972) have studied a self-scoring method (SSM) that provides immediate feedback on each response made by the student. The self-scoring method requires the student to continue responding to an item until the correct alternative is found. The test is scored by counting the number of responses required to answer all the items on the test.

The rationales for all of these grading procedures make the implicit assumption that the test items can discriminate between students who do not select the correct alternative as their first choice. This implies that the residual items that result from the elimination of one or more distractors have a positive discriminating power. This study investigates the effects on reliability when this assumption is violated. In particular, the effects of SSM grading on inter-item correlation are examined for items where the residual items contribute no discriminating power to the test. The inclusion of such items could explain the discouraging results that have often been obtained when procedures for awarding partial credit have been tried.

Results

Here we will consider some assumptions about the response probabilities for items. Let P_i be the probability that the i-th item is answered correctly on the first try, and let P_{ij} be the joint probability of a correct response on the first try for both the i-th and j-th items. If the first response is incorrect, assume that all subsequent responses result from random guessing. The probability of choosing the correct answer on any response, after the first, is $\frac{1}{n}$ where n is the number of alternatives left when the response is made. This model is the basis for the traditional formula scores that try to correct for guessing.

Let X_i and X_j be the student's scores on the i-th and j-th items, respectively. Every student must make at least one response to obtain the correct answer. Let R_i be the number of incorrect responses: $R_i=X_i-1$. The constant term does not affect the inter-item correlations, thus, $Cor(R_i,R_j)=Cor(X_i,X_j)$, where R_i and R_j have a joint discrete density function. The marginal distribution for the k-th item depends only on the probability of a correct response on the first choice, P_k, and the number of distractors:

$$P(R_k=0) = P_k$$
$$P(R_k=i) = (1-P_k)\left(\frac{1}{N-1}\right) \qquad 0 < i < N$$

The means and variances for a set of such items are presented below.

$$E(R_k) = (1-P_k)\frac{N}{2}$$
$$V(R_k) = (1-P_k)P_k\frac{N^2}{4}\left[1+\left(\frac{1}{P_k}\right)\left(\frac{N-2}{3N}\right)\right]$$

In order to compute the covariance, we need to know the joint density function for R_i and R_j. This distribution is given below.

$$P(R_i=0,R_j=0) = P_{ij}$$
$$P(R_i=0,R_j=J) = (P_i-P_{ij})\left(\frac{1}{N-1}\right)$$
$$P(R_i=I,R_j=0) = (P_j-P_{ij})\left(\frac{1}{N-1}\right)$$
$$P(R_i=I,R_j=J) = (1-P_i-P_j+P_{ij})\left(\frac{1}{N-1}\right)^2 \qquad 0 < I, J < N$$

The probability that $R_i=I\neq 0$ and $R_j=J\neq 0$ is just the probability that neither item is answered correctly on the first try, times the conditional probability that $R_i=I$ and $R_j=J$, given that neither item is answered correctly on the first try. So, for I and J greater than zero:

$$P(R_i=I,R_j=J) = P(R_i\neq 0,R_j\neq 0)\cdot P(R_i=I,R_j=J \mid R_i\neq 0,R_j\neq 0)$$
$$= (1-P_i-P_j+P_{ij})\frac{1}{(N-1)^2}.$$

The other probabilities are derived analogously.

Using the joint probability distribution presented above and the definition of covariance, we obtain:

$$Cov(R_i,R_j) = \frac{N^2}{4}(P_{ij}-P_iP_j).$$

The covariance for SSM grading of a given item is greater than the covariance for 0-1 grading by a factor of $\frac{N^2}{4}$. This increase results from the change in the range of possible scores from 0-1 to 1-(N-1).

Using the earlier results, the correlation between R_i and R_j is:

$$Cor(R_i,R_j) = \frac{(P_{ij}-P_iP_j)}{\sqrt{P_i(1-P_i)\left[1+\frac{1}{P_i}\left(\frac{N-2}{3N}\right)\right]\cdot P_j(1-P_j)\left[1+\frac{1}{P_j}\left(\frac{N-2}{3N}\right)\right]}}.$$

For reference, the conventional 0-1 grading system results in the following inter-item correlations:

$$\mathrm{Cor}(X_i, X_j) = \frac{P_{ij} - P_i P_j}{\sqrt{P_i(1-P_i)P_j(1-P_j)}} .$$

These two formulas differ only in the denominator, where the terms representing the variances of the two item scores are larger for SSM scoring than for conventional grading. In the case where any responses after the first response are at the chance level, SSM scoring will result in lower inter-item correlations than 0-1 scoring.

In changing from 0-1 to SSM grading, the covariance of two items and the variances for the individual items increase, but the proportional increases in the variances are greater than the increase in the covariances. When the distractors provide no discriminating power for students who cannot answer the question on their first try, SSM grading increases the proportion of the variance for any item that results from random guessing.

Discussion

The results derived above show that SSM grading does not necessarily increase inter-item correlations and therefore does not necessarily increase the reliability of a test. A test consisting exclusively of items with the properties hypothesized here could suffer a substantial decrease in reliability if SSM grading were substituted for conventional grading.

Although the quantitative effect of SSM grading on the reliability of multiple choice items has been examined here for a special case, some general qualitative conclusions can be drawn from the results. Gilman and Ferry (1972) hypothesized that the effect of SSM grading on reliability is equivalent to the inclusion of additional items. This is a useful way to look at the question, but it must be remembered that the new items will not have the same properties as the original items. Depending on the properties of these new items, SSM grading may either increase or decrease the reliability of the test, or leave it unchanged.

This paper demonstrated that the reliability of a test using SSM grading will depend on the properties of the item alternatives. The results developed here may explain the relatively small gains (in one case, a loss) in reliability that have often resulted from the use of procedures for awarding partial credit on multiple choice items.

If SSM is to be used to increase reliability, a currently recommended practice (see, for example, the discussion of distractors in Gronlund, 1965, p. 154) for the selection of distractors will have to be changed. Instead of choosing equally probable distractors, distractors should be generated whose probability of selection reflects the level of understanding of the student.

References

Gilman, D. A. and Ferry, P. Increasing test reliability through self-scoring procedures. Journal of Educational Measurement, 1972, 9, 205-207.

Hambleton, R. K., Roberts, D. M., and Traub, R. E. A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. Journal of Educational Measurement, 1970, 7, 75-82.

Gronlund, N. F. Measurement and Evaluation in Teaching. New York: MacMillan Company, 1968.

Koehler, R. A. A comparison of the validities of conventional choice testing and various confidence marking procedures. Journal of Educational Measurement, 1971, 8, 297-303.